Estimation of causal direction in the presence of latent confounders and linear non-Gaussian SEMs

Shohei Shimizu
Osaka University, Japan

with
Kenneth Bollen
University of North Carolina, Chapel Hill, USA
Abstract

• Estimation of causal direction of two observed variables in the presence of latent confounders
• A key challenge in causal discovery
• Propose a non-Gaussian method
• Not require to specify the number of latent confounders
• Experiments on artificial and sociology data
Background
Motivation

• Causality is a main interest in many empirical sciences

• Many recent methods for estimating causal directions (with no temporal information)
  – Linear non-Gaussian model (Dodge & Rousson 2001; Shimizu et al., 2006)
  – Nonlinear model (Hoyer et al., 2009; Zhang & Hyvarinen, 2009; Peters et al. 2011)

• Another important challenge: Latent confounders

![Diagram showing relationship between sleep problems and depression mood with a question mark indicating uncertainty about which is dominant.](Epidemiology (Rosenstrom et al., 2012))
Structural equation modeling (SEM) (Bollen, 1989; Pearl, 2000, 2009)

• A framework for describing causal relations
• An example (of linear cases):

\[ x_2 := f(x_1, e_2) = b_{21} x_1 + e_2 \]

– The value of \( x_2 \) is determined by the values of \( x_1 \) and error/exogenous variable \( e_2 \) through the linear function

• Generally speaking, if the value of \( x_1 \) is changed and that of \( x_2 \) also changes, then \( x_1 \) causes \( x_2 \)
Major challenges

1. Estimation of causal direction when temporal information is not available

2. Coping with latent confounders
Non-Gaussian approach: **LiNGAM**

*(Linear Non-Gaussian Acyclic Model)*  
(Shimizu et al., 2006)

- Acyclic SEMs with different directions **distinguishable**  
  (Dodge & Rousson, 2001; Shimizu et al., 2006)

\[
\begin{align*}
\text{Model 1:} & \quad x_1 &= e_1 \\
& \quad x_2 &= b_{21}x_1 + e_2
\end{align*}
\]

\[
\begin{align*}
\text{Model 2:} & \quad x_1 &= b_{12}x_2 + e_1 \\
& \quad x_2 &= e_2
\end{align*}
\]

where \(e_1\) and \(e_2\) are **error/exogenous** variables

- **Fundamental assumptions:**
  - \(e_1\) and \(e_2\) are non-Gaussian
  - Independence btw. \(e_1\) and \(e_2\) *(No latent confounders)*
Different directions give different data distributions

Model 1:
\[ x_1 = e_1 \]
\[ x_2 = 0.8x_1 + e_2 \]

Model 2:
\[ x_1 = 0.8x_2 + e_1 \]
\[ x_2 = e_2 \]

\[ E(e_1) = E(e_2) = 0, \]
\[ \text{var}(x_1) = \text{var}(x_2) = 1 \]
LiNGAM with latent confounders
(Hoyer, Shimizu & Kerminen, 2008)

- Extension to incorporate non-Gaussian latent confounders $f_q$

$$x_i = \mu_i + \sum_{q=1}^{Q} \lambda_{iq} f_q + \sum_{j\neq i} b_{ij} x_j + e_i$$

where, WLG, $f_q$ ($q = 1, \cdots, Q$) are independent:

$$x_1 = \mu_1 + \sum_{q=1}^{Q} \lambda_{1q} f_q + e_1$$

$$x_2 = \mu_2 + \sum_{q=1}^{Q} \lambda_{2q} f_q + b_{21} x_1 + e_2$$
Previous estimation approaches

• Explicitly model latent confounders and compare two models with opposite directions of causation
  – Maximum likelihood principle (Hoyer et al., 2008)
  – Bayesian model selection (Henao & Winther, 2011)
  – Laplace / finite mixture of Gaussians for $p(e_i)$

• Require to specify the number of latent confounders, which is difficult in general
Our proposal

Reference:
Shimizu and Bollen (2014)
Journal of Machine Learning Research
In press
Key idea (1/2)

• Another look at the LiNGAM with latent confounders:

\[ x^{(m)}_2 = \mu_2 + \sum_{q=1}^{Q} \lambda_{2q} f_q^{(m)} + b_{21} x^{(m)}_1 + e^{(m)}_2 \]

Observations are generated from the LiNGAM model with possibly different intercepts \( \mu_2 + \mu^{(m)}_2 \).
Key idea (2/2)

• Include the sums of latent confounders as the observation-specific intercepts:

\[ x^{(m)}_2 = \mu_2 + \sum_{q=1}^{Q} \lambda_{2q} f_q^{(m)} + b_{21} x_1^{(m)} + e_2^{(m)} \]

\( m \)-th obs.:

• Not explicitly model latent confounders
• Neither necessary to specify the number of latent confounders Q nor estimate the coefficients \( \lambda_{2q} \)
Our approach

- Compare these two LiNGAM models with opposite directions:

**Model 3 \((x_1 \rightarrow x_2)\)**

\[
\begin{align*}
x_1^{(m)} &= \mu_1 + \mu_1^{(m)} + e_1^{(m)} \\
x_2^{(m)} &= \mu_2 + \mu_2^{(m)} + b_{21}x_1^{(m)} + e_i^{(m)}
\end{align*}
\]

**Model 4 \((x_1 \leftarrow x_2)\)**

\[
\begin{align*}
x_1^{(m)} &= \mu_1 + \mu_1^{(m)} + b_{12}x_2^{(m)} + e_1^{(m)} \\
x_2^{(m)} &= \mu_2 + \mu_2^{(m)} + e_2^{(m)}
\end{align*}
\]

- Many additional parameters \(\mu_i^{(m)}\) \((i = 1, 2; m = 1, \ldots, n)\)
- Prior for the observation-specific intercepts \(\mu_i^{(m)}\)
- Other para. low-informative: Gaussian with large sd.
- Bayesian model selection (marginal likelihoods)
Prior for the observation-specific intercepts

\[ \mu_1^{(m)} = \sum_{q=1}^{Q} \lambda_{1q} f_q^{(m)}, \quad \mu_2^{(m)} = \sum_{q=1}^{Q} \lambda_{2q} f_q^{(m)} \]

- **Motivation:** Central limit theorem
  - Sums of independent variables tend to be more Gaussian
- **Approximate the density by a bell-shaped curve dist.**

\[
\begin{bmatrix}
\mu_1^{(m)} \\
\mu_2^{(m)}
\end{bmatrix}
\sim t\text{-distribution with sd } \sigma_1, \sigma_2 \text{, correlation } \sigma_{12}, \text{ and DOF } \nu
\]

- **Select the hyper-parameter values that maximize the marginal likelihood:** Empirical Bayes
  - \( \sigma_1 \in \{0, 0.2 \times \text{sd}(x_i), \ldots, 1.0 \times \text{sd}(x_i)\} \), \( \sigma_{12} \in \{0, \pm 0.1, \ldots, \pm 0.9\} \)
  - DOF \( \nu \) fixed to be 6 in the experiments below
- **Small \( \sigma_i \) means similar intercepts**
Experiments on artificial data
Experimental results (100 obs.)

- Data generated from LiNGAM with latent confounders
- Various non-Gaussian distributions
  - Laplace, Uniform, asymmetric dist. etc.
- Our method uses Laplace for $p(e_i)$

Numbers of successful discoveries (100 rep.)

N. latent confounders = 1

- Our mthd: 80
- Hoyer: 1, 4 conf.
- Henao: 1, 4, 10 conf.

N. latent confounders = 6

- Our mthd: 86
- Hoyer: 1, 4 conf.
- Henao: 1, 4, 10 conf.
Experiment on sociology data
Sociology data

- Source: General Social Survey (n=1380)
  - Non-farm background, ages 35-44, white, male, in the labor force, no missing data for any of the covariates, 1972-2006

Status attainment model
(Duncan et al., 1972)
Evaluation of our method using the sociology data

Known (temporal) orderings of 15 pairs

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>FO &lt;- FE</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>SI &lt;- FO</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>SI &lt;- FE</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>SI &lt;- SO</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>SI &lt;- SE</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>SI &lt;- NS</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>SO &lt;- FO</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>SO &lt;- SE</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>SO &lt;- NS</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>SE &lt;- FO</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>SE &lt;- FE</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>SE &lt;- NS</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>NS &lt;- FO</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>NS &lt;- FE</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

N. successes: 12 10 9 9 7 8
Precisions: 0.80 0.67 0.60 0.60 0.47 0.53
Conclusions
Conclusions

• Estimation of causal direction in the presence of latent confounders is a major challenge in causal discovery

• Our proposal: Fit linear non-Gaussian SEM with possibly different intercepts to data

• Future works
  – Test other informative priors for observation-specific intercepts
  – Implement a wider variety of error/prior distributions (e.g., learn DOF of t dist.)
  – Develop extensions using nonlinear/cyclic models (Hoyer et al., 2009; Zhang & Hyvarinen, 2009; Lacerda et al., 2008) instead of LiNGAM